Amostragem e medidas de qualidade de shapelets

Lucas S. Cavalcante

Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo

Orientador: Gustavo E. A. P. A. Batista

August 2, 2017

Estrutura

Definições & Notações

- Uma série temporal é uma sequência ordenada de valores reais de amostras obtidas em um intervalo fixo. Uma série de m amostras é denotada por $T = \{t_1, t_2, \ldots, t_m\},\ t_i \in \mathbb{R}$.
- A partir de uma série temporal é possvel extrair uma subsequência, que é uma sequência continua de amostras de uma série. Por exemplo, uma subsequência de 4 amostras com início na 5 amostra de T é denotada por $S = \{t_5, t_6, t_7, t_8\}$. Apesar disso, uma subsequência também é uma srie temporal.

Definições & Notações

Este trabalho está centrado na classificação supervisionada de séries temporais:

- Cada série temporal T possui um rótulo de classificação c, formando assim uma tupla $\langle T, c \rangle$.
- Por simplicidade este rótulo de classificação é denotado por um inteiro $1 \leq c \leq C$.

Um conjunto de n tuplas define um conjunto de dados $CD = \{\langle T_1, c_1 \rangle, \langle T_2, c_2 \rangle, \ldots, \langle T_n, c_n \rangle\}$.

Definições & Notações

Neste trabalho são feitas inúmeras comparações entre duas séries temporais, S e T, por meio de uma medida de dissimilaridade denotada por $dist(S, T)$.

- O conjunto de distâncias de uma série temporal S à todas as séries de um CD é denotado por $D_S = \{d_{S,1}, d_{S,2}, \ldots, d_{S,n}\}$.
- Como $d_{S,i} \in \mathbb{R}$ é possível denotar $\bar{D}_S$ como a média dessas medidas.

Por fim, denotaremos por $CD_{|c_i}$ e $D_{S|c_i}$ as partições desses conjuntos por uma classe c_i qualquer, e as suas cardinalidades por $|CD_{|c_i}|$ e $|D_{S|c_i}|$.

Aplicação e Classificação

O uso de técnicas de classificação de séries temporais tem sido aplicado em diversas áreas, por exemplo, em botânica foi aplicado à classificação de espécies [1], enquanto que na industria de manufatura foi aplicado para averiguar a qualidade de produção [2]; além de aplicações nas áreas de finanças, medicina e entretenimento [3].

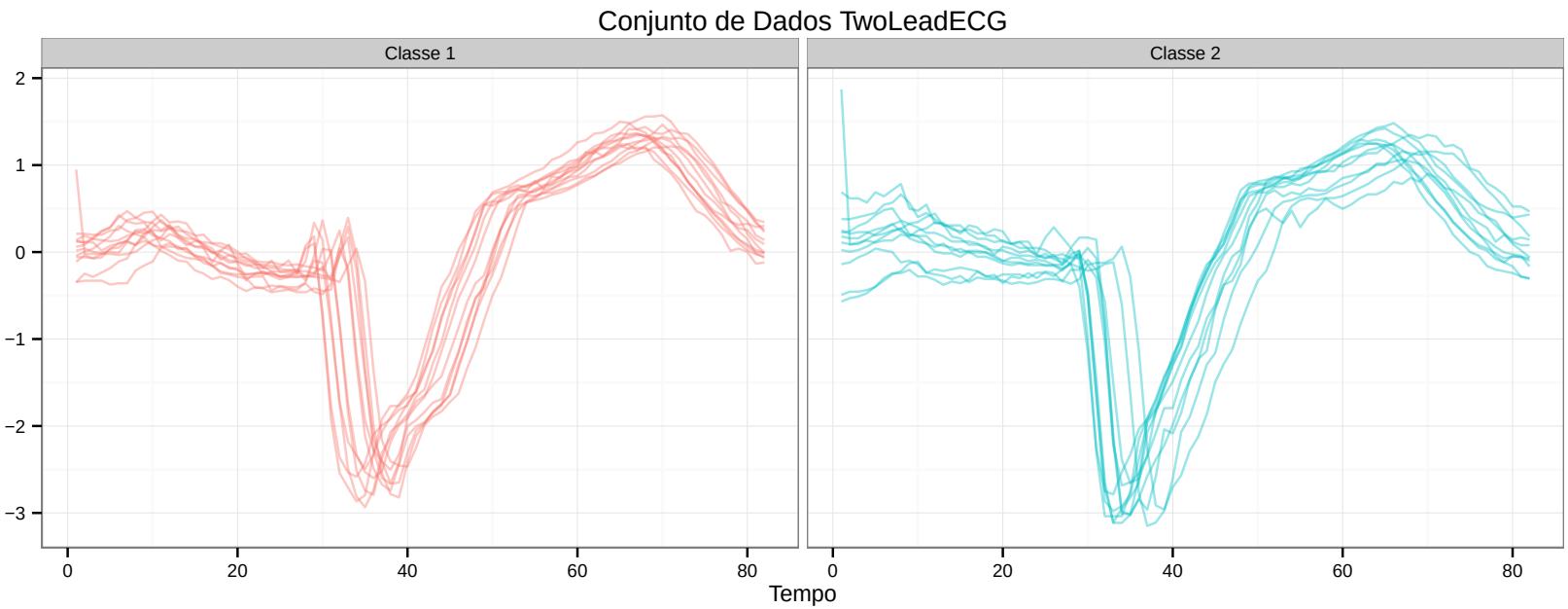
Classificação por kNN

Apesar da pluralidade de áreas de aplicação, a evidência empírica sugere que o simples classificador de vizinhos mais próximos (kNN) em conjunto com a medida de dissimilaridade *Dynamic Time Warping* (DTW) tem uma acurácia que é difícil de ser superada [4]. No entanto, esta abordagem tem pontos negativos [1]:

- Classificador de difícil compreensão;
- Custo elevado para classificar uma nova instância.

Shapelets

Shapelet é uma primitiva de séries temporais introduzida em 2009 que trata das limitações mencionadas. Esta primitiva é uma subsequência que de certa forma é representativa de sua classe [1].



Árvore de Decisão Shapelet

Esta capacidade do shapelet de representar uma classe advém do seu processo de extração e avaliação:

- Em um conjunto de n séries temporais, cada qual de tamanho m, existem $O(nm^2)$ subsequências. Cada uma delas é um candidato a bom shapelet.
- A avaliação dos shapelets se dá pelo seu ganho de informação, que exige a definição de uma medida de dissimilaridade entre uma subsequência e uma série temporal.

Originalmente esta primitiva foi embutida no classificador de árvore de decisão, no qual obteve alta acurácia, em especial em conjuntos de dados no qual a diferença entre as classes se dá por padrões locais [1].

Avaliação do Shapelet por Medidas de Qualidade

Intuitivamente, é questionável o uso do ganho de informação para avaliar um shapelet em um problema de múltiplas classes, dado que ele propõe uma partição binária do conjunto de dados. Por isso, se testou substituir o ganho de informação [5]:

- Ao se trocar o ganho de informação pelas medidas da Mediana de Mood e pelo teste não-paramétrico de Kruskal-Willis, não notou-se nenhuma diferença estatisticamente significativa entre as árvores de decisão em termos de acurácia, porém, ao se usar essas outras medidas foi possível reduzir o tempo de execução em até quase 20%.

Transformada Shapelet

Apesar da boa performance da árvore de decisão shapelet, um estudo recente mostrou que dissociar o processo de classificação do de extração de shapelets pode aumentar a acurácia [6].

- A dissociação é obtida por construir uma representação do conjunto de dados no qual shapelets se tornam atributos que quantificam numericamente uma série temporal. Esta é a transformada shapelet.
- Parte do aumento da acurácia advém da possibilidade de usar classificadores mais complexos do que uma árvore de decisão.

Transformada Shapelet

Ao se utilizar os classificadores simples de kNN, Árvore de decisão e Naive Bayes e os complexos SVM, Random Forest, Rotation Forest e Bayesian Networks sob a transformada notou-se que [6]:

- Em geral, classificadores simples tem uma acurácia pior do que os complexos, sendo que o SVM obteve a melhor acurácia geral.
- A conclusão é de que é preferível obter a transformada e empregar classificadores complexos, pois se obtém uma acurácia melhor e a interpretabildiade não é afetada (os atributos que são os shapelets continuam a serem interpretáveis).

Porém, notamos que esses experimentos foram realizados sob um espaço reduzido dos shapelets.

Re-Avaliação das Medidas de Qualidade

O trabalho que introduziu a transformada shapelet também fez uma re-avaliação das medidas de qualidade [6]:

- Além das medidas alternativas da Mediana de Mood e do teste não-paramétrico de Kruskal-Willis foi adicionada a *f-statistic* utilizada no teste ANOVA.
- Os experimentos mostraram que a árvore de decisão com a *f-statistic* possui a maior acurácia (sem significância estatística) e na média o menor tempo de execução. Com base nesses resultados os autores afirmaram: "*we argue that the f-statistic should be the default choice for shapelet quality*."

Nós notamos que os resultados deles vieram de experimentos sob a árvore de decisão shapelet, mas os autores fizeram tal afirmação para o caso da transformada shapelet também.

Lacunas & Objetivos

Utilizar todos os shapelets

Concordamos com a necessidade de se reduzir o espaço de busca por bons shapelets por questões de aplicações práticas. No entanto, é necessário ter experimentos que utilizam todos os shapelets para servir como *ground-truth*.

Utilização de Amostragem Aleatória

Propor o uso da amostragem aleatória para reduzir o espaço de busca por bons shapelets, ao invés de utilizar o algoritmo que restringe a busca por bons shapelets àqueles que tenham um determinado tamanho. Além disso, avaliar como essa redução afeta a acurácia.

Lacunas & Objetivos

Avaliação das Medidas de Qualidade

Fazer uma extensiva avaliação das medidas de qualidade sob o domínio da transformada shapelet. Além disso, introduzir uma nova medida de qualidade denominada *in-class transitions* e compará-la com as outras duas medidas abordadas neste trabalho: ganho de informação e *f-statistic*. A comparação será em relação a acurácia.

Medida de Dissimilaridade

A medida de Dissimilaridade adotada neste trabalho é a adaptação da Distância Euclidiana para o caso em que as séries temporais possuem diferentes quantidades de amostras [1]:

Distância Euclidiana entre duas Séries Temporais (S, T)

$$dist(S,T) = \min_{0 \leq \mathrm{j} < \mathrm{m}-\mathrm{m}'} \left\{ \sqrt{\frac{1}{m'} \sum_{i=1}^{m'} (s_i - t_{i+j})^2} \right\}$$

Com a observação de que sempre é realizada a z-normalização.

Shapelet

- Shapelet é uma subsequência que é representativa de uma classe.
- Inicialmente o shapelet era a subsequência de maior ganho de informação [1], mas com a transformada shapelet esta definição é relaxada para qualquer subsequência que tenha uma medida de qualidade associada a ela.
- Em um conjunto de dados existem $O(nm^2)$ shapelets.

Shapelet: Medida de Qualidade

Seja um shapelet Sh do qual se obteve D_{Sh}, então:

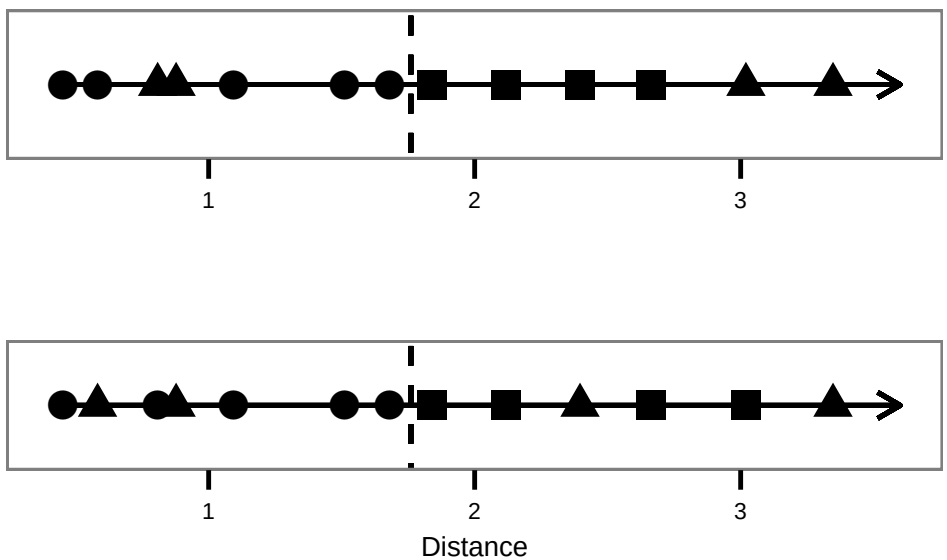
- **Ganho de Informação:** inicialmente se computa a entropia do conjunto de dados corrente, então por um corte binário se particiona este conjunto dois. O ganho de informação é a diferença entre a entropia inicial e a entropia média das partições resultantes [7].
- **F-Statistic:** é a mesma métrica utilizada no teste da ANOVA (*Analysis of Variance*) para testar a hipótese nula de que as médias de todas as classes são iguais.

Cálculo da F-Statistic

$$FStatistic(D_{Sh}) = \frac{\sum_{i=1}^{C}(\bar{D_{Sh|c_i}} - \bar{D_{Sh}})^2/(C-1)}{\sum_{i=1}^{C}\sum_{d_j \in D_{Sh|c_i}}(d_j - \bar{D_{Sh|c_i}})^2/(n-C)}$$

Shapelet: Medidas de Qualidade

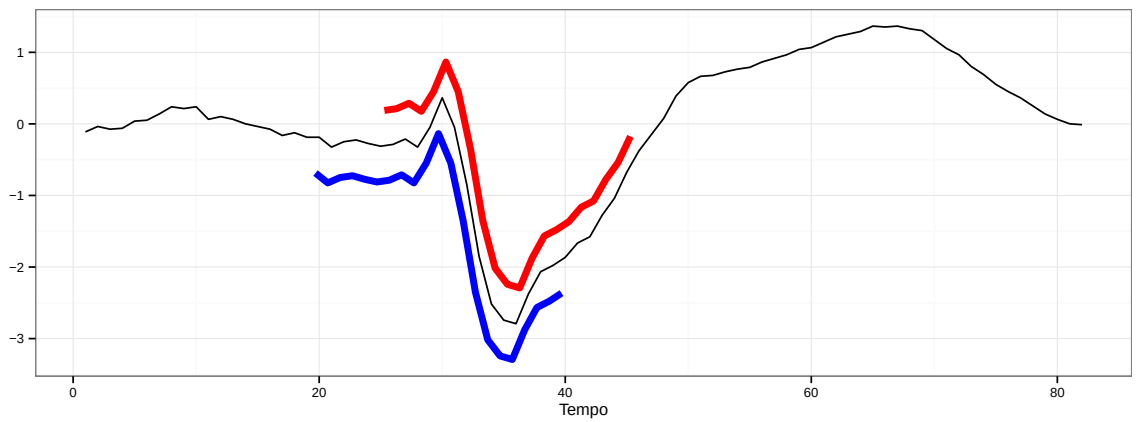
- **In-Class Transitions:** é a contagem de quantos elementos da mesma classe aparecem na ordenação de D_{Sh}.



Não importa a medida de qualidade, quanto maior seu valor melhor.

Shapelets: Seleção

Dado um conjunto de shapelets é preciso selecionar k deles, pois usar todos pode causar overfitting. Um método naive seria de simplesmente selecionar os k melhores, porém, isso pode levar a alta redundância de atributos.



Hills et. al [6] contorna esse problema por definir a auto-similaridade de shapelets.

Shapelets: Seleção

```
function SelecionarKShapelets(shapelets, k)
    selecionados ← ∅
    Ordenar(shapelets, "decrescente", "medida de qualidade")
    for shapelet ∈ shapelets do
        if NaoAutoSimilar(shapelet, selecionados) then
            selecionados ← selecionados ∪ shapelet
            k ← k − 1
        end if
        if k = 0 then
            return selecionados
        end if
    end for
    return selecionados
end function
```

Shapelets: Extração

- Extração de todos os $O(nm^2)$ shapelets.
- Extração de uma amostragem aleatória dos shapelets.
- Extração de todos os shapelets dentre *min* e *max* [6]:

function EstimarMinMax(*CD*)
    shapelets $\leftarrow \emptyset$
    for $i \leftarrow 1, 10$ **do**
        subset $\leftarrow$ SelecaoAleatoria(*CD*, 10)
        tmp $\leftarrow$ GerarShapelets(subset, 3, *m*)
        shapelets $\leftarrow$ shapelets $\cup$ SelecionrKShapelets(tmp, 10)
    end for
    Ordenar(shapelets, "crescente", "tamanho")
    min $\leftarrow$ shapelets[25]
    max $\leftarrow$ shapelets[75]
    return min, max
end function

Shapelets: Transformada

Após a extração dos shapelets e a compurtação de sua qualidade, então se escolhe k shapelets para compor o grupo de atributos Sh. Na transformada shapelet cada série temporal é representada pela sua distância aos shapelets em Sh.

Table: Descrição da transformada shapelet. Cada linha representa uma série temporal e cada coluna um shapelet, com uma coluna extra para o rótulo de classificação. Cada a_{ij} é a distância entre um shapelet é uma série temporal.

Time Series	Sh_1	Sh_2	$\ldots$	Sh_k	Class. Label
T_1	a_{11}	a_{12}	$\ldots$	a_{1k}	c_1
T_2	a_{21}	a_{22}	$\ldots$	a_{2k}	c_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
T_n	a_{n1}	a_{n2}	$\ldots$	a_{nk}	c_n

Shapelets: Classificação

Uma vez que se saiba obter a transformada shapelet de um *CD*, então para classificar novas séries temporais a partir de um conjunto de treinamento basta:

1. Obter a transformada shapelet do conjunto de treinamento.
2. Induzir um modelo de classificação sob tal transformada.
3. Obter a transformada shapelet do conjunto de teste.
4. Executar o modelo de classificação usando como entrada a transformada do conjunto de teste.

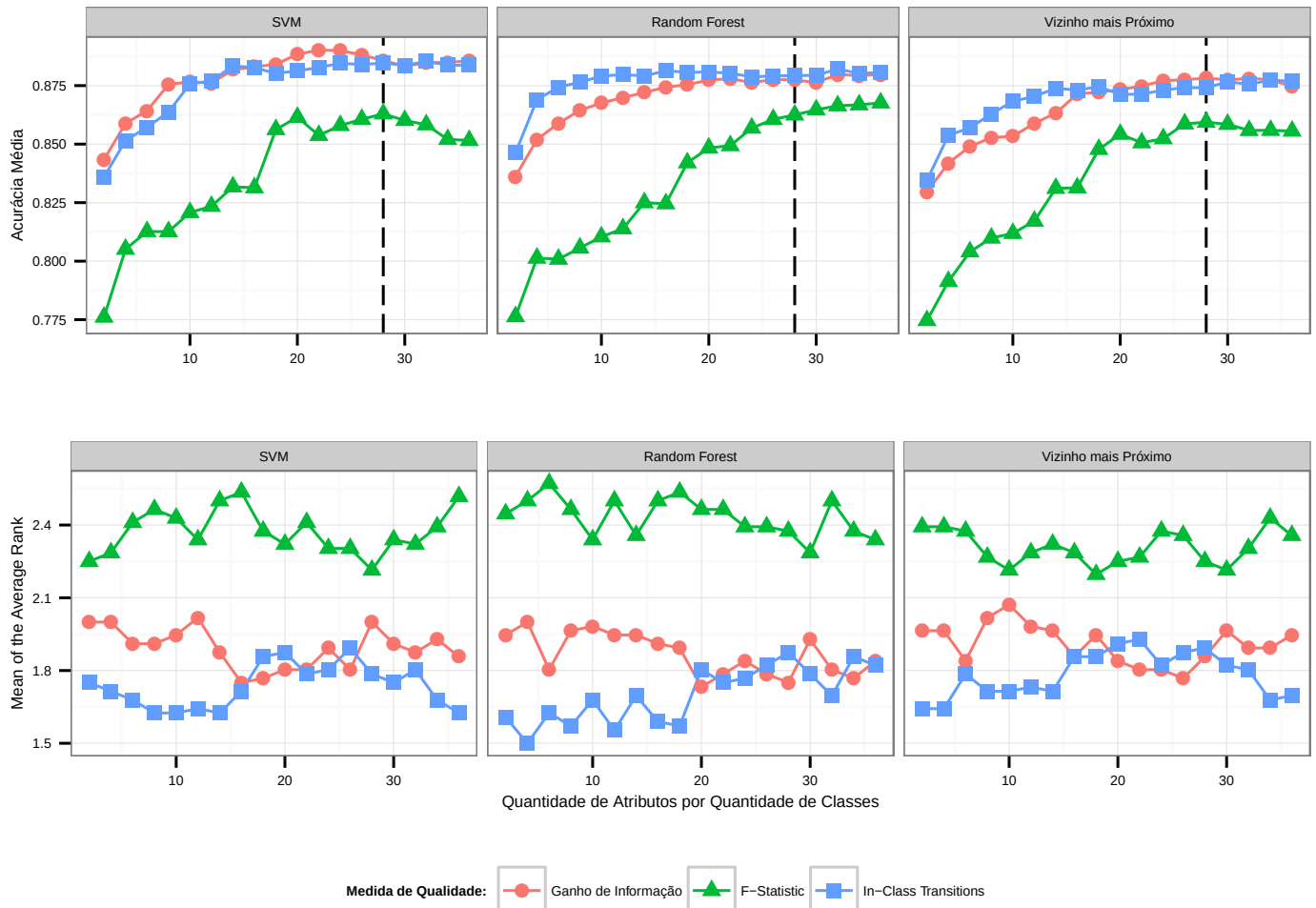
Neste trabalho serão explorados os seguintes classificadores: kNN, Random Forest e Support Vector Machines (SVM).

Arranjo Experimental

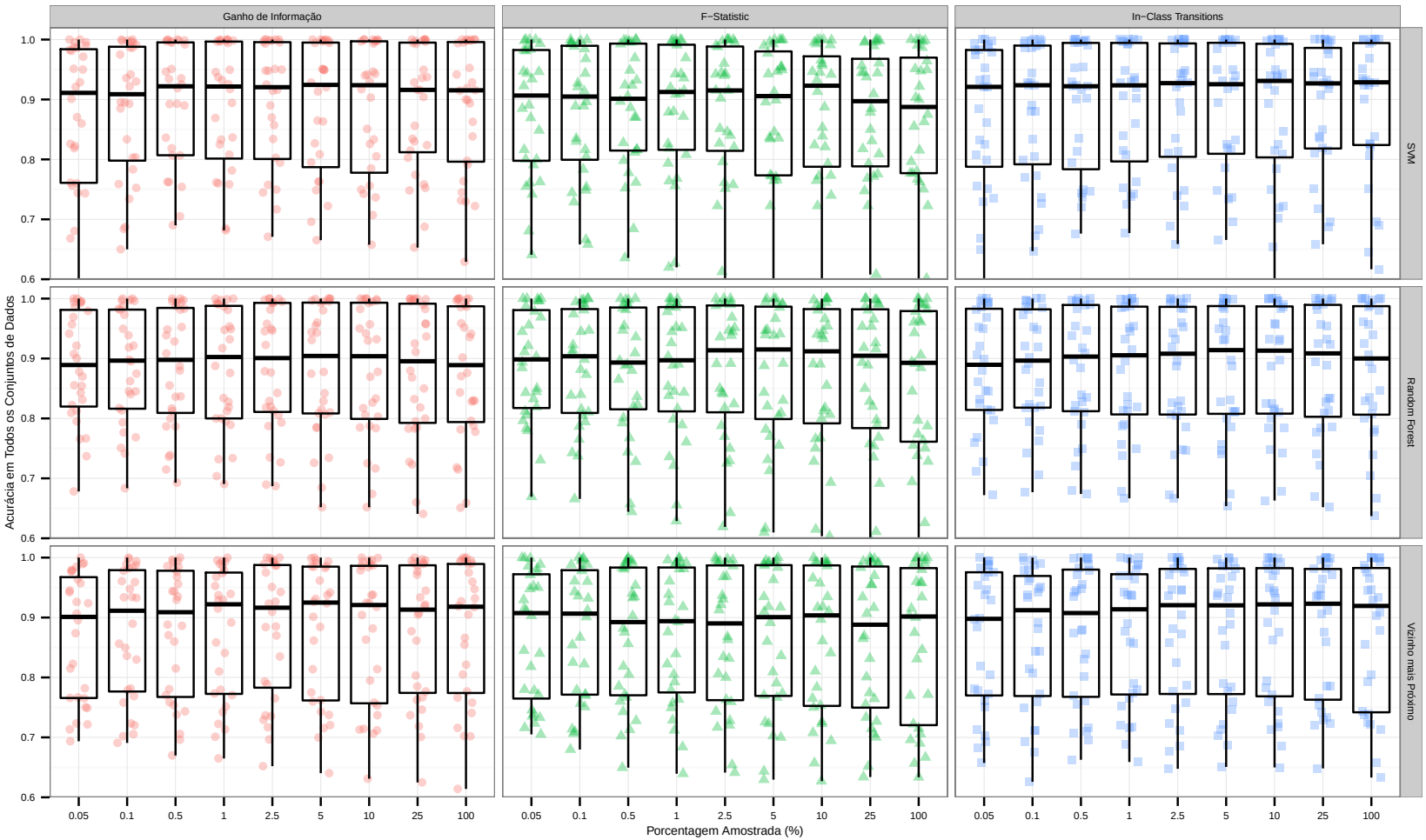
- Os experimentos foram executados em um Intel Core i7 @ 2.3GHz.
- A linguagem utilizada neste trabalho foi a R. Porém, para executar os experimentos em um tempo razoável, foi feita a paralelização e uso da linguagem C++ por meio do pacote RcppParallel.
- Todos os experimentos foram executados em uma coleção de 28 conjuntos de dados do repositório da UCR [8].

Com o detalhe de que a quantidade de atributos, em nossos experimentos, é dependente da quantidade de classes de cada conjunto de dados.

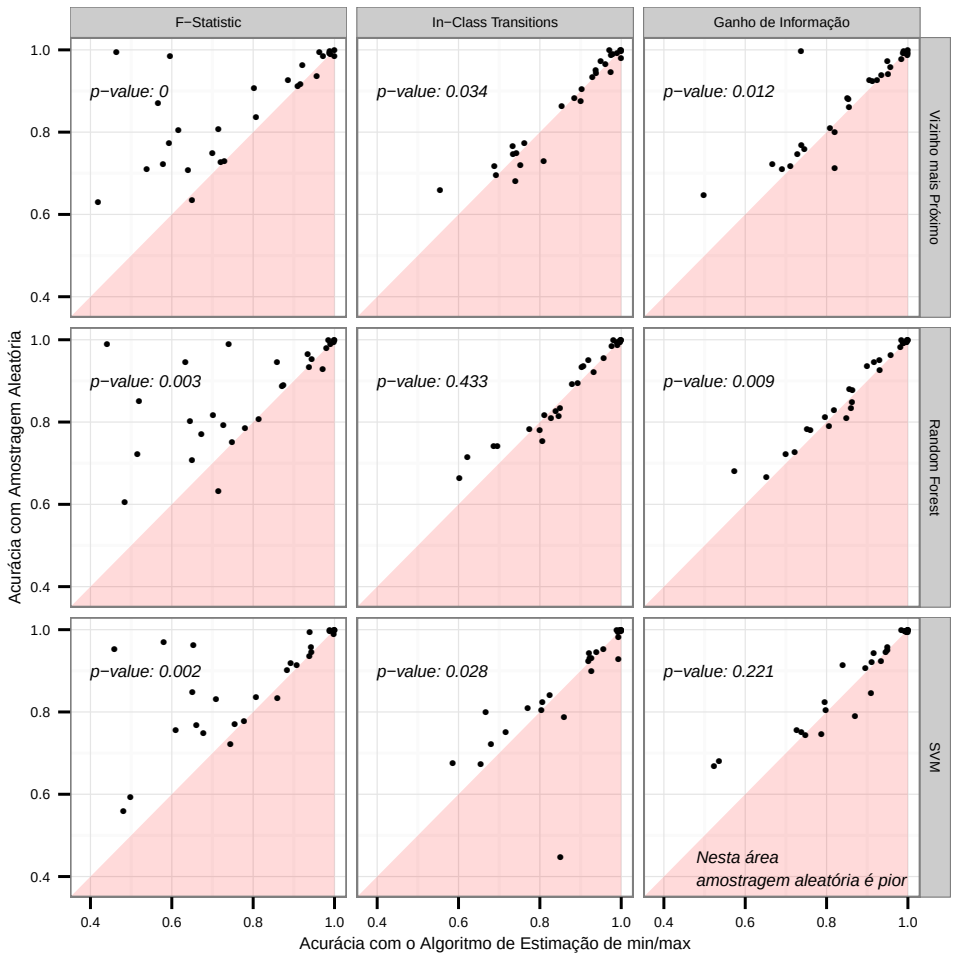
Avaliação de Acurácia das Medidas de Qualidade



Avaliação da Amostragem Aleatória



Comparativo entre as Técnicas de Redução do Espaço

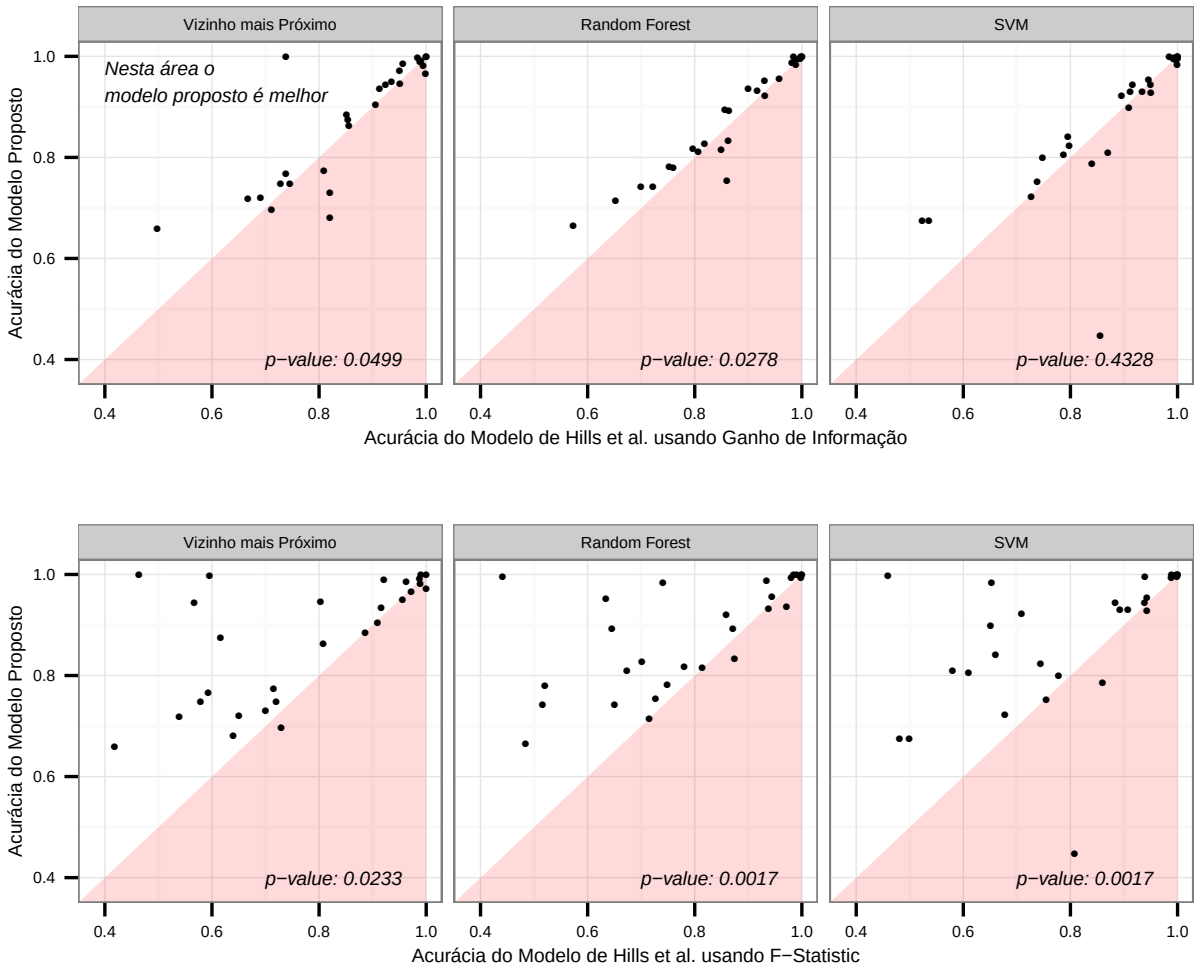


Comparativo entre as Técnicas de Redução do Espaço

Table: O algoritmo `EstimarMinMax()` estima uma faixa de valores de tamanho do qual todos os shapelets são extraídos. Dessa estimação calculamos o percentual que essa faixa representa do todo, então sumarizamos essa informação pelos quantis (5 execuções para cada conjunto de dados).

	Ganho de Informação	F-Stat	In-Class Transitions
Mínimo (%)	25,7	1,558	20,11
1^o Quartil (%)	31,51	10,56	31,3
Mediana (%)	35,31	16,08	36,52
Média (%)	37,24	18,28	36,28
3^o Quartil (%)	43,94	27,81	41,76
Máximo (%)	56,09	40,26	49,08

Comparação com o modelo de Hills et. al [6]



Conclusão & Trabalhos Futuros

- Em nossa análise sobre as medidas de qualidade os nossos achados contradizem a recomendação prévia do uso da f-statistic, pois ela obeteve a pior performance.
  - A nossa medida proposta, denominada in-class transitions, possui a melhor performance, especialmente quando poucos atributos são utilizados. Quando mais atributo são utilizados a sua performance se torna próxima a de ganho de informação.
  - Apesar da in-class transitions ser de fácil compreesão e de implementação ela ainda requer a ordenaç ao dos elementos, logo, deverá ser mais lenta que a f-statistic, porém mais rápida do que a de ganho de informação.

Conclusão & Trabalhos Futuros

- Como em um conjunto de dados existem $O(nm^2)$ shapelets, a computação de todos pode ser impraticável, logo métodos que reduzam o espaço de buscam se tornam uma necessidade.
  - O trabalho de Hills et al [6] sugere o uso de algoritmo `EstimarMinMax()` que restringe a busca por bons shapelets àqueles que tenham um determinado tamanho, enquanto que nós propusemos o uso de amostragem aleatória.
  - Nossos experimentos mostraram que o uso de `EstimarMinMax()` introduz um overhead significativo sem no entanto reduzir o espaço de busca de forma eficiente.
  - O uso de amostragem aleatória mostrou que pode oferecer um pequeno aumento da acurácia e diminuição de sua variabilidade. Supomos que isso de em razão de uma maior diversidade dos atributos.

Conclusão & Trabalhos Futuros

- Nós esperamos que as áreas de redução de espaço de busca e de prover maior diversidade nos shapelets recebam maior atenção no futuro. Por enquanto a nossa contribuiç ao se restringe a recomendar o uso da amostragem aleatória como o baseline para ambas as áreas.

Agradecimentos

Agradecimentos ao Daniel Y. T. Chino por sua sugestão de uso de amostragem aleatória para shapelets; e a FAPESP que financiou este trabalho (2013/16164-2)

Referências I

Lexiang Ye and Eamonn Keogh.

Time series shapelets: a new primitive for data mining.

In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 947–956. ACM, 2009.

Om P Patri, Abhishek B Sharma, Haifeng Chen, Guofei Jiang, Anand V Panangadan, and Viktor K Prasanna.

Extracting discriminative shapelets from heterogeneous sensor data.

In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 1095–1104. IEEE, 2014.

Referências II

Xiaopeng Xi, Eamonn Keogh, Christian Shelton, Li Wei and Chotirat Ann Ratanamahatana.

Fast time series classification using numerosity reduction.

In *Proceedings of the 23rd international conference on Machine learning*, pages 1033–1040. ACM, 2006.

Gustavo EAPA Batista, Eamonn J Keogh, Oben Moses Tataw, and Vinícius MA de Souza.

CID: an efficient complexity-invariant distance for time series.

Data Mining and Knowledge Discovery, 28(3):634–669, 2014.

Jason Lines and Anthony Bagnall.

Alternative quality measures for time series shapelets.

In *Intelligent Data Engineering and Automated Learning-IDEAL 2012*, pages 475–483. Springer, 2012.

Referências III

Jon Hills, Jason Lines, Edgaras Baranauskas, James Mapp, and Anthony Bagnall.

Classification of time series by shapelet transformation.

Data Mining and Knowledge Discovery, 28(4):851–881, 2014.

Ross J Quinlan.

C4. 5: Programs for machine learning.

1992.

Yanping Chen, Eamonn Keogh, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, and Gustavo Batista.

The ucr time series classification archive, July 2015.

www.cs.ucr.edu/~eamonn/time_series_data/.